

PRODUCT GUIDE 2026

# 製品ガイド

ひとつのトークンファクトリー、4つの使い方

One Token Factory. Four ways to use it.

01

## SIKUMO Cloud

GPUクラウド・GPU Cloud

02

## SIKUMO Models

オープンモデルAPI・Open Model API

03

## SIKUMO Factory

プライベート・トークンファクトリー

04

## SIKUMO Hub

モデルAPIハブ・Model API Hub



## OVERVIEW

# 製品全体像：ひとつのトークンファクトリー

SIKUMOのすべての製品は同じトークンファクトリー上で稼働し、国際標準のレイヤー構造で構築されています。最下層は日本・米国の法規リージョンに配置されたNVIDIA最新GPU（IaaS）、その上にクラウドネイティブなオーケストレーション、モデルサービング（MaaS）、AIゲートウェイのコントロールプレーン、最上層にお客様向けの4つの提供形態があります。



## 製品の選び方

学習や自社推論のためにGPUが必要

→ 01 SIKUMO Cloud

コスト効率の高いモデルを直接呼び出したい

→ 02 SIKUMO Models

データを外に出せない、専用のAI基盤が必要

→ 03 SIKUMO Factory

複数ベンダーのモデルを一元管理したい

→ 04 SIKUMO Hub

**オープンモデルを中核に、クローズドAPIを補完に：同等品質、より低いトークンコスト。** トークン単価 = コンピュートコスト ÷ (ハードウェア効率 × モデル効率 × 運用効率)

## 01 SIKUMO Cloud

### GPUクラウド・NVIDIA最新GPUを、GPUaaSとして提供

GPUの購入やデータセンター建設、計算環境の構築を行うことなく、学習・推論・ファインチューニング・AIアプリ運用に必要な計算資源を利用できます。リソースはSIKUMO自社のトークンファクトリーとパートナー算力から供給され、日本・米国などNVIDIA GPUサーバーを合規に調達できるリージョンに配置されています。

### コンピュータ仕様

#### NVIDIA B300

日本・学習＋推論

大規模モデルの学習、ファインチューニング、高スループット推論

#### NVIDIA Vera Rubin

米国・推論特化

大規模・低遅延のモデル推論とAPIサービス

### 提供形態

#### GPUインスタンス

オンデマンド構成。開発・検証・短期推論に

#### ベアメタル

物理サーバー占有。本番環境とモデル配備に

#### GPUクラスター

高速インターコネクト接続。分散学習と大規模推論に

#### 専用クラスター

大口のお客様向け専用リソースプール。長期安定を保証

### 利用・課金モデル

#### オンデマンド

ON-DEMAND

弾力的・短期タスク向け

#### リザーブド

RESERVED

月次／四半期コミットで  
コスト最適化

#### 年間契約

ANNUAL

長期安定の学習・本番運用

#### 専用

DEDICATED

クラスター丸ごと専用、  
カスタム提供

### こんなお客様に

AIスタートアップ・モデルチーム

学習・ファインチューニングが必要な企業

推論サービス・AIアプリ基盤

研究機関・大学

価格と開通方法はSIKUMOカスタマーチームまでお問い合わせください。提供リージョン・型番・納期は個別に確認します。

## 02 SIKUMO Models

### オープンモデルAPI・世界の優れたオープンモデルを、ひとつのAPIで、トークン課金

モデルの配備・運用を自前で行う必要はありません。SIKUMOは自社およびパートナーの算力上に世界の優良オープンモデルを配備し、推論最適化を施した標準APIとして提供、実際のトークン使用量で精算します。オープンソースコミュニティの最新版に継続的に追随します。

### モデルカタログ（随時更新）

#### DeepSeekシリーズ

推論・コード・汎用対話

コスト効率の高い最先端推論モデル

#### Qwen・GLM・Llama・Kimi・MiniMax など

多言語・長文コンテキスト・Agent

日本語・英語・中国語など多言語シーンに対応

### 提供形態

#### 標準API

主要API仕様と互換。既存アプリをスムーズに移行

#### ストリーミングとバッチ

オンラインのリアルタイム呼び出しとオフラインのバッチ処理

#### ツール呼び出し

Function Callingと構造化出力に対応（モデルによる）

#### 専用推論エンドポイント

スループットと遅延を占有保証。高並列業務に

### 利用・課金モデル

#### 従量

PAY-AS-YOU-GO

入力／出力トークンで課金

#### コミット

COMMITTED

使用量コミットで単価を低減

#### 専用エンドポイント

DEDICATED

GPU時間で課金

#### エンタープライズ

ENTERPRISE

カスタム価格とSLA

### こんなお客様に

AIアプリ・Agent開発者

SaaS・インターネット企業

多言語コンテンツ・カスタマーサポート

高価なクローズドAPIを置き換えたいチーム

価格と開通方法はSIKUMOカスタマーチームまで。モデルのバージョン・コンテキスト・能力範囲はモデルカードをご参照ください。

## 03 SIKUMO Factory

プライベート・トークンファクトリー・トークンファクトリーをまるごと、お客様の環境へ

金融・医療・公共・製造などデータ機微性の高い業界向けに、SIKUMOは算力、オープンモデル、推論基盤、ファインチューニング、運用をパッケージ化し、専用クラスター、オンプレミス、ハイブリッドで提供します。契約で定めたデータ境界、カスタマイズ可能なモデル、予測可能なコスト。

### 提供内容

#### 算力 + プラットフォーム

専用GPUクラスターと推論運用基盤  
コンピュータ基盤と推論基盤をベースに

#### モデル + ファインチューニング

オープンモデルのプライベート配備と業界特化調整  
企業の知識・用語・プロセスを組み込んだ専用能力

### 配備形態

#### 専用ホスティング

SIKUMOのデータセンター内に置く顧客専用クラスター

#### オンプレミス

お客様のデータセンターに配備、完全な自社管理

#### ハイブリッド

機微データはローカル処理、弾力的な負荷はクラウドへ

#### 継続運用

監視・アップグレード・モデル更新・運用サポート

### 導入プロセス

#### 評価

ASSESS

ユースケース・データ・コンプライアンス評価

#### 設計

DESIGN

算力とモデルの方案设计

#### 配備

DEPLOY

クラスター・モデル・基盤の稼働

#### 運用

OPERATE

ファインチューニング反復と継続運用

### こんなお客様に

金融機関

医療・ライフサイエンス

官公庁・公共部門

大手製造業

価格と開通方法はSIKUMOカスタマーチームまで。データ境界・権限・監査範囲は契約で明確化します。

## 04 SIKUMO Hub

### モデルAPIハブ・ひとつの接続で、提供可能なオープン／クローズドモデルを統合

複数のモデルベンダーと個別に接続する必要はありません。SIKUMO Hubは自社オープンモデルと認可済みの主要クローズドモデルAPIを統合し、統一ゲートウェイ・統一認証・統一請求を提供、コスト・品質・遅延・コンプライアンス要件に応じてスマートにルーティングします。

#### プラットフォーム機能

##### 統一接続

オープン＋クローズド、ひとつのインターフェース  
新モデルはコード変更なしで即利用可能

##### スマートルーティング

コスト・品質・遅延・コンプライアンス  
ポリシーに基づく最適モデル自動選択とフェイルオーバー

#### ガバナンス機能

##### 統一請求

全モデルの使用量をまとめて一枚の請求書に

##### 予算管理

チーム・プロジェクト・アプリ単位の上限とアラート

##### 使用量モニタリング

呼び出し数・遅延・エラー率・コストをリアルタイム表示

##### 監査ログ

企業のコンプライアンスと内部統制要件に対応

#### 利用の流れ

##### 接続

CONNECT

ひとつのAPIキーで全モデル  
に接続

##### ルーティング

ROUTE

ルーティングとフェイルオ  
ーバーポリシーを設定

##### ガバナンス

GOVERN

予算・権限・監査

##### 最適化

OPTIMIZE

コストと品質を継続的に最  
適化

#### こんなお客様に

マルチモデル活用企業

企業IT・AIプラットフォームチーム

AI Agent開発者

AI支出を一元管理したい組織

価格と開通方法はSIKUMOカスタマーチームまで。クローズドモデルは認可・リージョン・用途条件を満たす場合に限り接続します。

## INFRASTRUCTURE & TECHNOLOGY

# コンピュータ基盤と技術基盤

SIKUMOはNVIDIA GPUサーバーを合规に調達できるリージョンにのみトークンファクトリーを建設し、創業チームの10年以上にわたる分散コンピューティング運用経験を運用基盤としています。

JAPAN・TOKYO

## SIKUMO Tokyo

NVIDIA B300

学習＋推論・Training＋Inference

大規模モデルの学習・ファインチューニング・高スループット推論。日本とアジア太平洋のお客様へ

UNITED STATES

## SIKUMO US

NVIDIA Vera Rubin

推論特化・Inference-first

大規模・低遅延の推論とAPIサービス。北米とグローバルのお客様へ

## 4つの技術プラットフォーム（国際標準アーキテクチャ）

### コンピュータ基盤

Compute Platform・GPU IaaS、クラウドネイティブ編成とマルチテナント

### 推論基盤

Inference Platform・MaaS、推論・ファインチューニング・モデルレジストリ

### AIゲートウェイ

AI Gateway・統一API、ルーティング、課金、可観測性

### プライベートAIスタック

Private AI Stack・専用／オンプレ配備、セキュリティとコンプライアンス

## オープン標準、グローバルR&D協働

日本

Japan R&D

米国

US R&D

オープン標準

Open Standards

国際標準のアーキテクチャとオープン標準に従い、技術的中立性を保ちます。研究開発の成果はすべてのトークンファクトリーに継続的に反映され、世界中のお客様に届きます。

各ファクトリーの型番と段階は「計画／建設／検証／商用」で表記し、調達・納品・検収記録に基づきます。

SERVICE & SUPPORT

# サービスとサポート

## スタンダード

STANDARD

オンラインドキュメントとコンソール  
チケットサポート  
使用量・請求の照会

## エンタープライズ

ENTERPRISE

専任カスタマーマネージャー  
優先対応と技術サポート  
カスタム価格と契約条件

## デディケイテッド

DEDICATED

専任技術チーム  
配備・ファインチューニング・運用  
を一貫サポート  
専用SLA

## ご利用までの流れ

### 1 要件ヒアリング

ユースケース・規模・コンプライアンス要件

### 2 提案とお見積

製品構成と価格案

### 3 検証

テスト枠またはPoCの開通

### 4 本番稼働

契約・開通・継続サービス

## 私たちの約束

### 誠実

透明な価格、守れる約束、お客様とデータへの誠実さ

### 継続

インフラへの長期投資。10年頼れるパートナーに

### 創造

トークンコストを下げ続け、AIをよりシンプルで効果的に

## Superintelligence, at your service.

スーパーインテリジェンスを、あなたのために。

SIKUMO株式会社・SIKUMO K.K.

東京・Web [sikumo.ai](https://sikumo.ai)・Email \_\_\_\_\_・Tel \_\_\_\_\_